

# On learning parametric-output HMMs

Aryeh Kontorovich

Department of Computer Science  
Ben-Gurion University  
Beer Sheva, Israel

Boaz Nadler

Department of Computer Science and applied Mathematics  
Weizmann Institute of Science  
Rehovot, Israel

Roi Weiss

Department of Computer Science  
Ben-Gurion University  
Beer Sheva, Israel

## Abstract

We present a novel approach to learning an HMM whose outputs are distributed according to a parametric family. This is done by *de-coupling* the learning task into two steps: first estimating the output parameters, and then estimating the hidden states transition probabilities. The first step is accomplished by fitting a mixture model to the output stationary distribution. Given the parameters of this mixture model, the second step is formulated as the solution of an easily solvable convex quadratic program. We provide an error analysis for the estimated transition probabilities and show they are robust to small perturbations in the estimates of the mixture parameters. Finally, we support our analysis with some encouraging empirical results.

## 1 Introduction

Hidden Markov Models (HMM) are a standard tool in the modeling and analysis of time series with a wide variety of applications. When the number of hidden states is known, the standard method for estimating the HMM parameters from given observed data is the Baum-Welch algorithm [Baum et al., 1970]. The latter is known to suffer from two serious drawbacks: it

tends to converge (i) very slowly and (ii) only to a local maximum. Indeed, the problem of recovering the parameters of a general HMM is provably hard, in several distinct senses [Abe and Warmuth, 1992, Lyngsø and Pedersen, 2001, Terwijn, 2002].

In this paper we consider learning parametric-output HMMs with a finite and known number of hidden states, where the output from each hidden state follows a parametric distribution from a given family. A notable example is a Gaussian HMM, where from each state  $x$ , the output is a (possibly multivariate) Gaussian,  $\mathcal{N}(\mu_x, \Sigma_x)$ , typically with unknown  $\mu_x, \Sigma_x$ .

**Main results.** We propose a novel approach to learning parametric output HMMs, based on the following two insights: (i) in an ergodic HMM, the stationary distribution is a mixture of distributions from the parametric family, and (ii) given the output parameters, or their approximate values, one can efficiently recover the corresponding transition probabilities up to small additive error.

Combining these two insights leads to our *decoupling* approach to learning parametric HMMs. Rather than attempting, as in the Baum-Welch algorithm, to jointly estimate both the transition probabilities and the output density parameters, we instead learn each of them separately. First, given one or several long observed sequences, the HMM output parameters are estimated by a general purpose parametric mixture learner, such as the Expectation-Maximization (EM) algorithm. Next, once these parameters are approximately known, we learn the hidden state transition probabilities by solving a computationally efficient convex quadratic program (QP).

The key idea behind our approach is to treat the underlying hidden process as if it were sampled independently from the Markov chain’s stationary distribution, and operate only on the empirical distribution of singletons and consecutive pairs. Thus we avoid computing the exact likelihood, which depends on the full sequence, and obtain considerable gains in computational efficiency. Under mild assumptions on the Markov chain and on its output probabilities, we prove in Theorem 1 that given the exact output probabilities, our estimator for the hidden state transition matrix is asymptotically consistent. Additionally, this estimator is robust to small perturbations in the output probabilities (Theorems 2-6).

Beyond its practical prospects, our proposed approach also sheds light on the theoretical difficulty of the full HMM learning problem: It shows that for parametric-output HMMs the *key difficulty is fitting a mixture model*, since once its parameters have been accurately estimated, learning the transition

matrix can be cast as a convex program. While learning a general mixture is considered a hard problem, we note that recently much progress has been made under various separation conditions on the mixture components, see e.g. Moitra and Valiant [2010], Belkin and Sinha [2010] and references therein.

**Related work.** The problem of estimating HMM parameters from observations has been actively studied since the 1970's, see Cappé et al. [2005], Rabiner [1990], Roweis and Ghahramani [1999]. While computing the maximum-likelihood estimator for an HMM is in general computationally intractable, under mild conditions, such an estimator is asymptotically consistent and normally distributed, see Bickel et al. [1998], Chang [1996], Douc and Matias [2001].

In recent years, there has been a renewed interest in learning HMMs, in particular under various assumptions that render the learning problem tractable [Faragó and Lugosi, 1989, Hsu et al., 2009, Mossel and Roch, 2006, Siddiqi et al., 2010, Anandkumar et al., 2012]. Also, Cybenko and Crespi [2011], Lakshminarayanan and Raich [2010] recently suggested Non-negative Matrix Factorization (NNMF) approaches for learning HMMs. These methods are related to our approach, since with known output probabilities, NNMF reduces to a convex program similar to the one considered here. Hence, our stability and consistency analysis may be relevant to NNMF-based approaches as well.

**Paper outline.** In Section 2 we present our problem setup. The algorithm for learning the HMM appears in Section 3, and its statistical analysis in Section 4. Section 5 contains some simulation results. The technical details are deferred to the Appendices.

## 2 Problem Setup

**Notation.** When  $X \in \mathcal{X}$  and  $Y \in \mathcal{Y}$  take values in a discrete set we abbreviate  $P(x)$  for  $\Pr(X = x)$  and  $P(y|x)$  for  $\Pr(Y = y | X = x)$ . When  $Y \in \mathcal{Y}$  is continuous-valued, we denote by  $P(y|x)$  the probability density function of  $Y$  given  $X$ .

For  $x, w \in \mathbb{R}^n$ ,  $\text{diag}(x)$  denotes the  $n \times n$  diagonal matrix with entries  $x_i$  on its diagonal,  $x/w$  is the vector with entries  $x_i/w_i$ , and  $\|x\|_w^2 = \sum_i w_i x_i^2$  is a  $w$ -weighted  $\ell_2$  norm (for  $w_i > 0$ ). The shorthand  $x \lesssim y$  means  $x \leq$

$(1 + o(1))y$ . Similarly we write  $x \lesssim_P y$  for  $x \leq (1 + o_P(1))y$ . Finally, for a positive integer  $n \in \mathbb{N}$ , we write  $[n] = \{1, 2, \dots, n\}$ .

**Hidden Markov Model.** We consider a discrete-time, discrete-space HMMs with  $n$  hidden states. The HMM output alphabet, denoted  $\mathcal{Y}$ , may be either discrete or continuous. A parametric-output HMM is characterized by a tuple  $(A, \mathcal{F}_n^\theta, P_0)$  where  $A$  is an  $n \times n$  column stochastic matrix,  $P_0$  is the distribution of the initial state and  $\mathcal{F}_n^\theta = (f_{\theta_1}, \dots, f_{\theta_n})$  is an ordered tuple of *parametrized* probability density functions. In the sequel we sometimes write  $f_i$  instead of  $f_{\theta_i}$ .

To generate the output sequence of the HMM, first an unobserved Markov sequence of hidden states  $x = \{x_t\}_{t=0}^{T-1}$  is generated with the following distribution.

$$P(x) = P_0(x_0) \prod_{t=1}^{T-1} A_{x_t, x_{t-1}},$$

where  $A_{ij} = P(X_t = i | X_{t-1} = j)$  are the transition probabilities. Then, each hidden state  $X_t$  independently emits an observation  $Y_t \in \mathcal{Y}$  according to the distribution  $P(y_t | x_t) \equiv f_{x_t}(y_t)$ . Hence the output sequence  $y = (y_t)_{t=0}^{T-1}$  has the conditional probability

$$P(y | x) = \prod_{t=0}^{T-1} P(y_t | x_t) = \prod_{t=0}^{T-1} f_{x_t}(y_t).$$

**The HMM Learning Problem.** Given one or several HMM output sequences  $(Y_t)_{t=0}^{T-1}$ , the HMM learning problem is to estimate both the transition matrix  $A$  and the parameters of the output distributions  $\mathcal{F}_n^\theta$ .

### 3 Learning Parametric-Output HMMs

The standard approach to learning the parameters of an HMM is to maximize the likelihood

$$\sum_{x \in [n]^T} P_0(x_0) P(y_0 | x_0) \prod_{t=1}^{T-1} A_{x_t, x_{t-1}} P(y_t | x_t).$$

As discussed in the Introduction, this problem is in general computationally hard. In practice, neglecting the small effect of the initial distribution  $P_0(x_0)$

on the likelihood,  $A$  and  $\mathcal{F}_n^\theta$  are usually estimated via the Baum-Welch algorithm, which is computationally slow and only guaranteed to converge to a local maximum.

### 3.1 A Decoupling Approach

In what follows we show that when the output distributions are parametric, we can *decouple* the HMM learning task into two steps: learning the output parameters  $\theta_1, \dots, \theta_n$  followed by learning the transition probabilities of the HMM. Under some mild structural assumptions on the HMM, this decoupling implies that the difficulty of learning a parametric-output HMM can be reduced to that of learning a parametric mixture model. Indeed, given (an approximation to)  $\mathcal{F}_n^\theta$ 's parameters, we propose an efficient, single-pass, statistically-consistent algorithm for estimating the transition matrix  $A$ .

As an example, consider learning a Gaussian HMM with univariate outputs. While the Baum-Welch approach jointly estimates  $n^2 + 2n$  parameters (the matrix  $A$  and the parameters  $\mu_i, \sigma_i^2$ ), our decoupling approach first fits a mixture model with only  $3n$  parameters  $(\pi_i, \mu_i, \sigma_i^2)$ , and then solves a convex problem for the matrix  $A$ . While both problems are in general computationally hard, ours has a significantly lower dimensionality for large  $n$ .

**Assumptions.** To recover the matrix  $A$  and the output parameters  $\theta_j$  we make the following assumptions:

(1a) The Markov chain has a unique stationary distribution  $\boldsymbol{\pi}$  over the  $n$  hidden states. Moreover, each hidden state is recurrent with a frequency bounded away from zero:  $\min_k \pi_k \geq a_0$  for some constant  $a_0 > 0$ .

(1b) The  $n \times n$  transition matrix  $A$  is geometrically ergodic<sup>1</sup>: there exists parameters  $G < \infty$  and  $\psi \in [0, 1)$  such that from any initial distribution  $P_0$

$$\|A^t P_0 - \boldsymbol{\pi}\|_1 \leq 2G\psi^t, \quad \forall t \in \mathbb{N}. \quad (1)$$

(1c) The output parameters of the  $n$  states are all distinct:  $\theta_i \neq \theta_j$  for  $i \neq j$ . In addition, the parametric family is identifiable.

**Remarks:** Assumption (1a) rules out transient states, whose presence makes it generally impossible to estimate all entries in  $A$  from one or a few long observed sequences. Assumption (1b) implies mixing and is used later on to bound the error and the number of samples needed to learn the

---

<sup>1</sup>Any finite-state ergodic Markov chain is geometrically ergodic.

matrix  $A$ . Assumption (1c) is crucial to our approach, which uses the distribution of only single and pairs of consecutive observations. If two states  $i, j$  had same output parameters, it would be impossible to distinguish between them based on single outputs.

### 3.2 Learning the output parameters.

Assumptions (1a,1b) imply that the Markov chain over the hidden states is mixing, and so after only a few time steps, the distribution of  $X_t$  is very close to stationary. Assuming for simplicity that already  $X_0$  is sampled from the stationary distribution, or alternatively neglecting the first few outputs, this implies that each observable  $Y_t$  is a random realization from the following *parametric mixture model*,

$$Y \sim \sum_{i=1}^n \pi_i f_{\theta_i}(y). \quad (2)$$

Hence, given the output sequence  $(Y_t)_{t=0}^{T-1}$  one may estimate the output parameters  $\theta_i$  and the stationary distribution  $\pi_i$  by fitting a mixture model of the form (2) to the observations. This is commonly done via the EM algorithm.

Like its more sophisticated cousin Baum-Welch, the mixture-learning EM algorithm also suffers from local maxima. Indeed, from a theoretical viewpoint, learning such a mixture model (i.e. the parameters of  $\mathcal{F}_n^\theta$ ) is a non-trivial task considered in general to be computationally hard. Nonetheless, under various separation assumptions, efficient algorithms with rigorous guarantees have been recently proposed (see e.g. Belkin and Sinha [2010]).<sup>2</sup> Note that while these algorithms have polynomial complexity in sample size and output dimension, they are still exponential in the number of mixture components (i.e., in the number of hidden states of the HMM). Hence, these methods do not imply polynomial learnability of parametric-output HMMs.

In what follows we assume that using some mixture-learning procedure, the output parameters  $\theta_j$  have been estimated with a relatively small error (say  $|\hat{\theta}_j - \theta_j| = O(1/\sqrt{T})$ ). Furthermore, to allow for cases where  $\theta_j$  were estimated from separate observed sequences of perhaps other HMMs with same output parameters but potentially different stationary distributions, we do not assume that  $\pi_i$  have been estimated.

---

<sup>2</sup>Note that the techniques for learning mixtures assume iid data. However, if these are algorithmically stable — as such methods typically are — the iid assumption can be replaced by strong mixing [Mohri and Rostamizadeh, 2010].

### 3.3 Learning the transition matrix $A$

Next, we describe how to recover the matrix  $A$  given either exact or approximate knowledge of the HMM output probabilities. For clarity and completeness, we first give an estimation procedure for the stationary distribution  $\pi$ .

**Discrete observations.** As a warm-up to the case of continuous outputs, we start with HMMs with a discrete observation space of size  $|\mathcal{Y}| = m$ . In this case we can replace  $\mathcal{F}_n^\theta$  by an  $m \times n$  column-stochastic matrix  $B$  such that  $B_{ki} \equiv P(k|i)$  is the probability of observing an output  $k$  given that the Markov chain is in hidden state  $i$ . In what follows, we assume that the number of output states is larger or equal to the number of hidden states,  $m \geq n$ , and that the  $m \times n$  matrix  $B$  has full rank  $n$ . The latter is the discrete analogue of assumption (1c) mentioned above.

First note that since the matrix  $A$  has a stationary distribution  $\pi$ , the process  $Y_t$  also has a stationary distribution  $\rho$ , which by analogy to Eq. (2), is

$$\rho = B\pi. \quad (3)$$

Similarly, the pair  $(Y_t, Y_{t+1})$  has a unique stationary distribution  $\sigma$ , given by

$$\sigma_{k,k'} = \sum_{\ell, \ell' \in [n]} \pi_\ell A_{\ell', \ell} B_{k, \ell} B_{k', \ell'}. \quad (4)$$

As we shall see below, knowledge of  $\rho$  and  $\sigma$  suffices to estimate  $\pi$  and  $A$ . Although  $\rho$  and  $\sigma$  are themselves unknown, they are easily estimated from a single pass on the data  $(Y_t)_{t=0}^{T-1}$ :

$$\begin{aligned} \hat{\rho}_k &= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{1}_{\{y_t=k\}}, \\ \hat{\sigma}_{k,k'} &= \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{1}_{\{y_{t-1}=k\}} \mathbb{1}_{\{y_t=k'\}}. \end{aligned} \quad (5)$$

**Estimating the stationary distribution  $\pi$ .** The key idea in our approach is to replace the exact, but complicated and non-convex likelihood function by a “pseudo-likelihood”, which treats the hidden state sequence  $(X_t)$  as if they were iid draws from the unknown stationary distribution

$\boldsymbol{\pi}$ . The pseudo-likelihood has the advantage of having an easily computed global maximum, which, as we show in Section 4, yields an asymptotically consistent estimator. Approximating the  $(X_t)$  as iid draws from  $\boldsymbol{\pi}$  means that the  $(Y_t)$  are treated as iid draws from  $\boldsymbol{\rho} = B\boldsymbol{\pi}$ . Thus, given a sequence  $(Y_t)_{t=0}^{T-1}$  the pseudo-likelihood for a vector  $\boldsymbol{\pi}$  is

$$\mathcal{L}(y_0, \dots, y_{T-1} \mid \boldsymbol{\pi}) = \prod_{i=0}^{T-1} (B\boldsymbol{\pi})_{y_i} = \prod_{k=1}^m (B\boldsymbol{\pi})_k^{n_k}$$

where  $n_k = \sum_{i=0}^{T-1} \mathbb{1}_{\{y_t=k\}} = T\hat{\rho}_k$ . Its maximizer is

$$\hat{\boldsymbol{\pi}}^{\text{ML}} = \arg \min_{x_i \geq 0, \|\mathbf{x}\|_1=1} - \sum_{k=1}^m \hat{\rho}_k \log(Bx)_k. \quad (6)$$

Since  $-\log(x)$  is convex,  $(Bx)_k$  is a linear combination of the unknown variables  $x_j$ , and the constraints are all linear, the above is nothing but a *convex* program, easily solved via standard optimization methods [Nesterov and Nemirovskii, 1994].

However, to facilitate the analysis and to increase the computational efficiency, we consider the asymptotic behavior of the pseudo-likelihood in (6), for  $T$  sufficiently large so that  $\hat{\boldsymbol{\rho}}$  is close to  $\boldsymbol{\rho}$ . First, we write

$$(Bx)_k = \hat{\rho}_k \left( 1 + \frac{(Bx)_k - \hat{\rho}_k}{\hat{\rho}_k} \right).$$

Next, assuming that  $T \gg 1$  is sufficiently large to ensure  $|(Bx)_k - \hat{\rho}_k| \ll \hat{\rho}_k$ , we take a second order Taylor expansion of  $\log(Bx)_k$  in (6). This gives

$$\begin{aligned} & - \sum_{k=1}^n \hat{\rho}_k \log \hat{\rho}_k - \sum_{k=1}^n ((Bx)_k - \hat{\rho}_k) + \\ & + \sum_{k=1}^n \hat{\rho}_k \left( \frac{(Bx)_k - \hat{\rho}_k}{\hat{\rho}_k} \right)^2 + O \left( \frac{\|Bx - \hat{\boldsymbol{\rho}}\|_\infty^3}{\min_j \rho_j^2} \right). \end{aligned}$$

The first term is independent of  $x$ , whereas the second term vanishes. Thus, we may approximate (6) by the *quadratic program*

$$\arg \min_{x_i \geq 0, \|\mathbf{x}\|_1=1} \|\hat{\boldsymbol{\rho}} - Bx\|_{(1/\hat{\boldsymbol{\rho}})}^2 \quad (7)$$

where  $\|x\|_w^2 = \sum_k w_k x_k^2$  is a weighted  $\ell_2$  norm w.r.t. the weight vector  $w$ . Eq. (7) is also a convex problem, easily solved via standard optimization techniques. However, let us temporarily ignore the non-negativity constraints  $x_i \geq 0$  and add a Lagrange multiplier for the equality constraint



$\sum x_i = 1$ :

$$\min \frac{1}{2} \sum_{k=1}^m \frac{1}{\hat{\rho}_k} \left( \hat{\rho}_k - \sum_{j=1}^n B_{kj} x_j \right)^2 - \lambda \left( \sum_j x_j - 1 \right). \quad (8)$$

Differentiating with respect to  $x_i$  yields

$$Wx = (1 + \lambda)\mathbf{1}, \quad (9)$$

where  $W = B^\top \text{diag}(1/\hat{\rho})B$ . Enforcing the normalization constraint is equivalent to solving for  $x^* = W^{-1}\mathbf{1}$  and normalizing  $\hat{\pi} = x^*/\|x^*\|_1$ . Note that if all entries of  $x^*$  are positive,  $\hat{\pi}$  is the solution of the optimization problem in (7), and we need not invoke a QP solver. Assumptions (1a,1b) that  $\pi_k$  is bounded away from zero and that the chain is mixing imply that for sufficiently large  $T$ , all entries of  $\hat{\pi}$  will be positive with high probability, see Section 4.

**Estimating the transition matrix  $A$ .** To estimate  $A$ , we consider pairs  $(Y_t, Y_{t+1})$  of consecutive observations. By definition we have that for a single pair,

$$P(Y_t = k, Y_{t+1} = k') = \sum_{i,j} B_{k'i} B_{kj} A_{ij} P(X_t = j).$$

As above, we treat the  $T - 1$  consecutive pairs  $(Y_t, Y_{t+1})$  as independent of each other, with the hidden state  $X_t$  sampled from the stationary distribution  $\pi$ . When the output probability matrix  $B$  and the stationary distribution  $\pi$  are both known, the pseudo-likelihood is given by

$$\mathcal{L}(y|A) = \prod_{(k,k')} \left( \sum_{ij} B_{k'i} B_{kj} A_{ij} \pi_j \right)^{n_{kk'}},$$

where  $n_{kk'} = \sum_{t=1}^{T-1} \mathbb{1}_{\{y_t=k\}} \mathbb{1}_{\{y_{t+1}=k'\}} = (T-1)\hat{\sigma}_{kk'}$ . The resulting estimator is

$$\underset{A_{ij} \geq 0, \sum_i A_{ij} = 1, A\pi = \pi}{\text{argmin}} - \sum \hat{\sigma}_{kk'} \log \left( \sum_{ij} C_{ij}^{kk'} A_{ij} \right) \quad (10)$$

where  $C_{ij}^{kk'} = \pi_j B_{kj} B_{k'i}$ . In practice, since  $\pi$  is not known, we use  $\hat{C}_{ij}^{kk'} = \hat{\pi}_j B_{kj} B_{k'i}$ , with  $\hat{\pi}$  instead of  $\pi$ . Again, (10) is a convex program in  $A$  and may be solved by standard constrained convex optimization methods. To obtain a more computationally efficient formulation, let us assume that  $\min_{k,k'} \sigma_{k,k'} \geq a_2 > 0$ , and that  $\min_{k,k'} T \hat{\sigma}_{kk'} \gg 1$ , so that

$|(\hat{C}A)_{kk'} - \hat{\sigma}_{kk'}| \ll \hat{\sigma}_{kk'}$ , where  $(\hat{C}A)_{kk'} = \sum_{ij} \hat{C}_{ij}^{kk'} A_{ij}$ . Then, as above, the approximate minimization problem is

$$\underset{A_{ij} \geq 0, \sum_i A_{ij} = 1, A\hat{\pi} = \hat{\pi}}{\operatorname{argmin}} \left\| \hat{\sigma} - \hat{C}A \right\|_{1/\hat{\sigma}}^2. \quad (11)$$

In contrast to the estimation of  $\pi$ , where we could ignore the non-negativity constraints, here the constraints  $A_{ij} \geq 0$  are essential, since for realistic HMMs, some entries in  $A$  might be strictly zero. Finally, note that if  $\hat{\pi} = \pi$  and  $\hat{\sigma} = \sigma$ , the true matrix  $A$  satisfies  $\sigma = CA$  and is the minimizer of (10).

In summary, given one or more output sequences  $(y_t)_{t=0}^{T-1}$  and an estimate of  $B$ , we first make a single pass over the data and construct the estimators  $\hat{\rho}$  and  $\hat{\sigma}$ , with complexity  $O(T)$ . Then, the stationary distribution  $\pi$  is estimated via (9), and its transition matrix  $A$  via (11). To estimate  $A$ , we first compute the matrix product  $\hat{C}^\top \hat{C}$ , with  $O(n^4 m^2)$  operations. The resulting QP has size  $n^2$ , and is thus solvable [den Hertog, 1994] in time  $O(n^6)$  — which is dominated by  $O(n^4 m^2)$  since  $m \geq n$  by assumption. Hence, the overall time complexity of estimating  $A$  is  $O(T + n^4 m^2)$ .

**Extension to continuous observations.** We now extend the above results to the case of continuous outputs distributed according to a known parametric family. Recall that in this case, each hidden state  $i \in [n]$  has an associated output probability density  $f_{\theta_i}(y)$ . As with discrete observations, we assume that an approximation  $(\hat{\theta}_1, \dots, \hat{\theta}_n)$  to  $f_i$ 's parameters is given and use it to construct estimates of  $\pi$  and  $A$ .

To this end, we seek analogues of (3) and (4), which relate the observable quantities to the latent ones. This will enable us to construct the appropriate empirical estimates and the corresponding quadratic programs, whose solutions will be our estimators  $\hat{\pi}$  and  $\hat{A}$ . To handle infinite output alphabets, we map each observation  $y$  to an  $n$ -dimensional vector  $\varphi(y) = (f_{\theta_1}(y), \dots, f_{\theta_n}(y))$ , whose entries are the likelihood of  $y$  from each of the underlying hidden states. As shown below, this allows us to reduce the problem to a discrete “observation” space which can be solved by the methods introduced in the previous subsection.

**Estimating the stationary distribution  $\pi$ .** To obtain an analogue of (3), we define the vector  $\xi \in \mathbb{R}^n$ , and matrix  $K \in \mathbb{R}^{n \times n}$ , which will play the role of  $\rho$  and  $B$  for discrete output alphabets. The vector  $\xi$  is defined as

$\xi = \mathbf{E}[\varphi(Y)]$ , or more explicitly,

$$\xi_k \equiv \mathbf{E}[f_k(Y)] = \sum_{j=1}^n \pi_j \int_{\mathcal{Y}} f_k(y) P(y | j) dy.$$

Similarly, the  $(i, j)$  entry of  $K$  is given by

$$K_{ij} \equiv \mathbf{E}[f_i(Y) | X = j] = \int_{\mathcal{Y}} f_i(y) P(y | j) dy. \quad (12)$$

With these definitions we have, as in Eq. (3),

$$\xi = K\pi. \quad (13)$$

Thus, given an observed sequence  $(y_t)_{t=0}^{T-1}$  we construct the empirical estimate

$$\hat{\xi}_k = \frac{1}{T} \sum_{t=0}^{T-1} f_k(y_t), \quad (14)$$

and consequently solve the QP

$$\hat{\pi} = \underset{\|x\|_1=1, x \geq 0}{\operatorname{argmin}} \left\| \hat{\xi} - Kx \right\|_{1/\hat{\xi}}^2. \quad (15)$$

In analogy to the discrete case, we assume  $\operatorname{rank}(K) = n$  so (15) has a unique solution. Its asymptotic consistency and accuracy are discussed in Section 4.

**Estimating the transition matrix  $A$ .** Next, following the same paradigm we obtain an analogue of (4). Bayes rule implies that for stationary chains,

$$P(k | Y) = \frac{f_k(Y) \pi_k}{\sum_{l=1}^n f_l(Y) \pi_l}. \quad (16)$$

We define the matrices  $\eta \in \mathbb{R}^{n \times n}$  and  $F \in \mathbb{R}^{n \times n}$  (analogues of  $\sigma$  and  $B$ ) as follows. Let  $Y$  and  $Y'$  be two *consecutive* observations of the HMM, then

$$\begin{aligned} \eta_{kk'} &\equiv \mathbf{E} [P(k | Y) P(k' | Y')] \\ F_{kj} &\equiv \mathbf{E} [P(k | Y) | j] = \int_{\mathcal{Y}} P(k | y) P(y | j) dy. \end{aligned} \quad (17)$$

A simple calculation shows that, as in (4),

$$\eta_{kk'} = \sum_{i,j=1}^n F_{k'i} F_{kj} A_{ij} \pi_j. \quad (18)$$

Since here  $F$  plays the role of  $B$ , we may call it an *effective* observation matrix. This suggests estimating  $A$  with the same tools used in the discrete case. Thus, given an observed sequence  $(y_t)_{t=0}^{T-1}$  we construct an empirical estimate  $\hat{\eta}$  by

$$\hat{\eta}_{kk'} = \frac{1}{T-1} \sum_{t=1}^{T-1} \hat{P}(k | y_{t-1}) \hat{P}(k' | y_t), \quad (19)$$

where  $\hat{P}$  is given by (16) but with  $\pi$  replaced by  $\hat{\pi}$ . Consequently we solve the following QP

$$\hat{A} = \underset{A_{ij} \geq 0, \sum_i A_{ij}=1, A\hat{\pi}=\hat{\pi}}{\operatorname{argmin}} \left\| \hat{\eta} - (\hat{C}A) \right\|_{1/\hat{\eta}}^2, \quad (20)$$

where  $\hat{C}_{ij}^{kk'} = \hat{\pi}_j F_{kj} F_{k'i}$  and  $(\hat{C}A)_{kk'} = \sum_{ij} \hat{C}_{ij}^{kk'} A_{ij}$ . As for the matrix  $B$  in the discrete case, to ensure a unique solution to Eq. (20) we assume  $\operatorname{rank}(F) = n$ .

**Remark 1.** *Instead of (18), we could estimate  $\eta'_{k,k'} \equiv \mathbf{E}[f_k(Y)f_{k'}(Y')]$ , from which  $A$  can also be recovered, since*

$$\eta'_{k,k'} = \sum_{i,j=1}^n K_{k'i} K_{kj} A_{ij} \pi_j.$$

*This has the advantage that for many distributions the matrix  $K$  can be cast in a closed analytic form. For example in the Gaussian case, while  $F$  needs to be calculated numerically, we have*

$$K_{ij} = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_i^2 + \sigma_j^2}} \exp \left( -\frac{1}{2} \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} \right).$$

*Additionally,  $K$  does not depend on the stationary distribution. The drawback is that in principle, and as simulations suggest, accurately estimating  $\eta'$  may require many more samples, see Appendix for details.*

In summary, given approximate output parameters  $(\hat{\theta}_1, \dots, \hat{\theta}_n)$ , we first calculate the  $n \times n$  matrix  $K$ . Next, we construct the vector  $\hat{\xi}$  by a single pass over the data  $(Y_t)_{t=0}^{T-1}$ . Then the stationary distribution  $\pi$  is estimated via (15). Given  $\hat{\pi}$ , we calculate the  $n \times n$  matrix  $F$ , construct the empirical estimate  $\hat{\eta}$ , and estimate  $A$  via (20). As in the discrete observation case, the time complexity of this scheme is  $O(T + n^6)$  with additional terms for calculating  $K$  and  $F$ .

## 4 Error analysis

First, we study the statistical properties of our estimators under the assumption that the output parameters,  $(\theta_1, \dots, \theta_n)$  in the continuous case, or the matrix  $B$  in the discrete case, are known *exactly*. Later on we show that our estimators are stable to perturbations in these parameters. For simplicity, throughout this section we assume that the initial hidden state  $X_0$  is sampled from the stationary distribution  $\pi$ . This assumption is not essential and omitting it would not qualitatively change our results. All proofs are deferred to the Appendices.

To provide bounds on the error and required sample size we make the following additional assumptions:

- (2a) In the discrete case, there exists an  $a_1 > 0$  such that  $\min_j \rho_j \geq a_1$ .
- (2b) In the continuous case, all  $f_{\theta_i}$  are bounded:

$$\max_{i \in [n]} \sup_{y \in \mathbb{R}} f_{\theta_i}(y) \leq L < \infty.$$

Finally, for ease of notation we define

$$g_\psi \equiv \frac{2G}{1 - \psi}.$$

**Asymptotic Strong Consistency.** Our first result shows that with perfectly known output probabilities, as  $T \rightarrow \infty$ , our estimates  $\hat{\pi}, \hat{A}$  are strongly consistent.

**Theorem 1.** *Let  $(Y_t)_{t=0}^{T-1}$  be an observed sequence of an HMM, whose Markov chain satisfies Assumptions (1a,1b). Assume  $\text{rank}(B) = n$  in the discrete case, or  $\text{rank}(F) = \text{rank}(K) = n$  in the continuous case. Then, both estimators,  $\hat{\pi}$  of (9) and  $\hat{A}$  of (11) in the discrete case, or (15) and (20) in the continuous case, are asymptotically strongly consistent. Namely, as  $T \rightarrow \infty$ , with probability one,*

$$\hat{\pi} \rightarrow \pi \quad \text{and} \quad \hat{A} \rightarrow A.$$

**Error analysis for the stationary distribution  $\pi$ .** Recall that to estimate  $\pi$  in the discrete case, we argued that for sufficiently large sample size  $T$ , the positivity constraints can be ignored, which amounts to solving an  $n \times n$  system of linear equations, Eq. (9). The following theorem provides both a lower bound on the required sample size  $T$  for this condition to hold with high probability, as well as error bounds on the difference  $\hat{\pi} - \pi$ .

**Theorem 2. Discrete case:** Let  $\hat{\rho}$  be given by (5), and  $\hat{\pi}$  be the solution of (9). Let  $\tilde{B} = \text{diag}(1/\sqrt{\hat{\rho}})B$ , and  $\sigma_1(\tilde{B})$  be its smallest singular value. Under Assumption (2a), a sequence of length

$$T \gtrsim \frac{g_\psi \sqrt{\log n}}{a_0 a_1 \sigma_1(\tilde{B})}, \quad (21)$$

is sufficient to ensure that with high probability, all entries in  $\hat{\pi}$  are strictly positive. Furthermore, as  $T \rightarrow \infty$ ,

$$\|\hat{\pi} - \pi\|_2 \lesssim_P \sqrt{\frac{g_\psi^2}{T a_1^2 \sigma_1^2(\tilde{B})}}. \quad (22)$$

Next we consider the errors in the estimate  $\hat{\pi}$  for the continuous observations case. For simplicity, instead of analyzing the quadratic program (15) with a weighted  $\ell_2$  norm, we consider the following quadratic program, whose solution is also asymptotically consistent:

$$\min_{x \geq 0, \sum_i x_i = 1} \|\hat{\xi} - Kx\|_2^2. \quad (23)$$

This allows for a cleaner analysis, without changing the qualitative flavor of the results.

**Theorem 3. Continuous case:** Let  $\hat{\xi}$  be given by (14),  $\hat{\pi}$  be the solution of (15), and  $\tilde{K} = \text{diag}(1/\sqrt{\hat{\xi}})K$ . Under Assumption (2b), as  $T \rightarrow \infty$ ,

$$\|\hat{\pi} - \pi\|_2 \lesssim_P \sqrt{\frac{(n^3 \ln n) g_\psi^2 L^4}{T \sigma_1^4(\tilde{K})}}, \quad (24)$$

**Error Analysis for the Matrix  $A$ .** Again, for simplicity, instead of analyzing the quadratic programs (11) and (20) with a weighted  $\ell_2$  norm, we consider the following quadratic programs, whose solutions are also asymptotically consistent for  $\hat{\nu} \in \{\hat{\sigma}, \hat{\eta}\}$ :

$$\min_{A_{ij} \geq 0, \sum_i A_{ij} = 1} \|\hat{\nu} - \hat{C}A\|_2^2. \quad (25)$$

Note that this QP is applicable even if  $\nu_{kk'} = 0$  for some  $k, k'$ , which implies that  $\hat{\nu}_{kk'} = 0$  as well.

**Theorem 4. Discrete case.** *Let  $\hat{A}$  be the solution of (25) with  $\hat{\nu} = \hat{\sigma}$  given in (5). Then, as  $T \rightarrow \infty$ ,*

$$\|\hat{A} - A\|_F \lesssim_P \sqrt{\frac{n^3 g_\psi^2}{T a_0^4 a_1^2 \sigma_1^{10}(B)}} \quad (26)$$

*and thus an observed sequence length*

$$T \gtrsim \frac{n^3 g_\psi^2}{a_0^4 a_1^2 \sigma_1^{10}(B)} \quad (27)$$

*suffices for accurate estimation.*

**Theorem 5. Continuous case.** *Let  $\hat{A}$  be the solution of (25) with  $\hat{\nu} = \hat{\eta}$  given in (19). Then, as  $T \rightarrow \infty$ ,*

$$\|\hat{A} - A\|_F \lesssim_P \sqrt{\frac{(n^7 \ln n) g_\psi^2 L^4}{T a_0^6 \sigma_1^8(F) \sigma_1^4(K)}} \quad (28)$$

*and thus an observed sequence length*

$$T \gtrsim \frac{(n^7 \ln n) g_\psi^2 L^4}{a_0^4 \sigma_1^8(F) \sigma_1^4(K)} \quad (29)$$

*suffices for accurate estimation.*

**Remarks.** Note the key role of the smallest singular value  $\sigma_1$ , in the error bounds in the theorems above: Two hidden states with very similar output probabilities drive  $\sigma_1$  to zero, thus requiring many more observations to resolve the properties of the underlying hidden sequence.

**Inaccuracies in the output parameters.** In practice we only have approximate output parameters, found for example, via an EM algorithm. For simplicity, we study the effect of such inaccuracies only in the continuous case. Similar results hold in the discrete case. To this end, assume the errors in the matrices  $K$  and  $F$  of Eqs. (12) and (17) are of the form

$$\tilde{K} = K + \epsilon LQ, \quad \tilde{F} = F + \epsilon P, \quad (30)$$

with  $\|Q\|_F, \|P\|_F \leq 1$ . The following theorem shows our estimators are *stable* w.r.t. errors in the estimated output parameters. Note that if  $K, F$  are estimated by a sequence of length  $T$ , then typically  $\epsilon = O(T^{-1/2})$ .

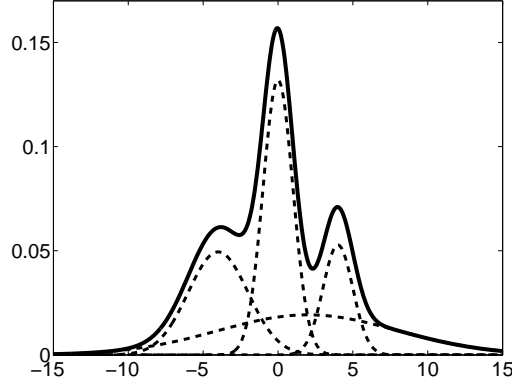


Figure 1: The mixture and its components.

**Theorem 6.** *Given an error of  $\epsilon$  in the output parameters as in Eq. (30), the estimators given in Theorems 3 and 5, incur an additional error of at most*

$$O\left(\frac{n^r \epsilon}{a_0^2 \sigma_1^4}\right), \quad (31)$$

with  $r = 1$  for estimating  $\pi$ , and  $r = \frac{3}{2}$  for estimating  $A$ , and where  $\sigma_1$  is the smallest singular value of  $K/L^2$  when estimating  $\pi$ , and of  $F$  when estimating  $A$ .

## 5 Simulation Results

We illustrate our algorithm by some simulation results, executed in MATLAB with the help of the HMM and EM toolboxes<sup>3</sup>. We consider a toy example with  $n = 4$  hidden states, whose outputs are univariate Gaussians,  $\mathcal{N}(\mu_i, \sigma_i^2)$ , with  $A$ ,  $\mathcal{F}_n^\theta$  and  $\pi$  given by

$$A = \begin{pmatrix} 0.7 & 0.0 & 0.2 & 0.5 \\ 0.2 & 0.6 & 0.2 & 0.0 \\ 0.1 & 0.2 & 0.6 & 0.0 \\ 0.0 & 0.2 & 0.0 & 0.5 \end{pmatrix}, \quad \begin{aligned} f_1 &= \mathcal{N}(-4, 4) \\ f_2 &= \mathcal{N}(0, 1) \\ f_3 &= \mathcal{N}(2, 36) \\ f_4 &= \mathcal{N}(4, 1) \end{aligned}$$

$$\pi^\top = (0.3529, 0.2941, 0.2353, 0.1176).$$

Fig. 1 shows the mixture and its four components.

<sup>3</sup>Available at <http://www.cs.ubc.ca/~murphyk> and <http://www.mathworks.com/> (under EM\_GM\_Fast).



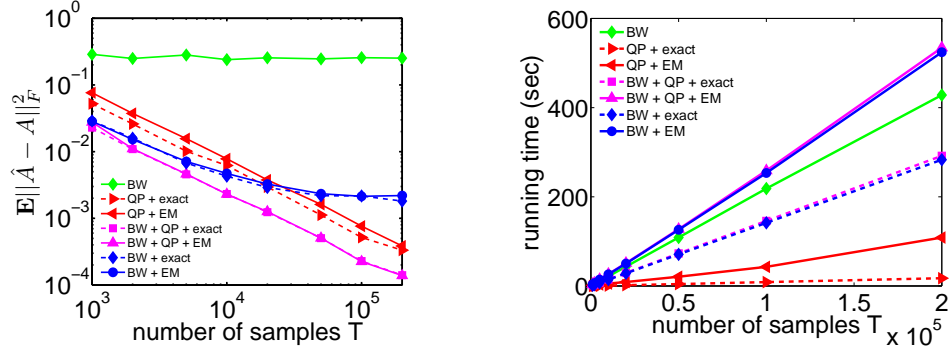


Figure 2: Average Error  $\mathbf{E}\|\hat{A} - A\|_F^2$  and runtime comparison of different algorithms vs. sample size  $T$ .

To estimate  $A$  we considered the following methods:

	method	initial $\theta$	initial $A$
1	BW	random	random
2	none	exactly known	QP
3	none	EM	QP
4	BW	exactly known	QP
5	BW	EM	QP
6	BW	exactly known	random
7	BW	EM	random

Fig. 2 (left) shows on a logarithmic scale  $\mathbf{E}\|\hat{A} - A\|_F^2$  vs. sample size  $T$ , averaged over 100 independent realizations. Fig. 2 (right) shows the running time as a function of  $T$ . In these two figures, the number of iterations of the BW step was set to 20.

Fig. 3 (left) shows the convergence of  $\mathbf{E}\|\hat{A} - A\|_F^2$  as a function of the number of BW iterations, with known output parameters, but either with or without the QP results. Fig. 3 (right) gives  $\mathbf{E}\|\hat{A} - A\|_F^2$  as a function of the number of BW iterations for both known and EM-estimated output parameters with  $10^5$  samples.

The simulation results highlight the following points: (i) BW with a random guess of both  $A$  and the parameters  $\theta_j = (\mu_j, \sigma_j^2)$  is useless if run for only 20 iterations. It often requires hundreds of iterations to converge, in some cases to a poor inaccurate solution (results not shown due to lack of space); (ii) For a small number of samples the accuracy of QP+EM (method 3) is comparable to BW+EM (method 5) but requires only a fraction of

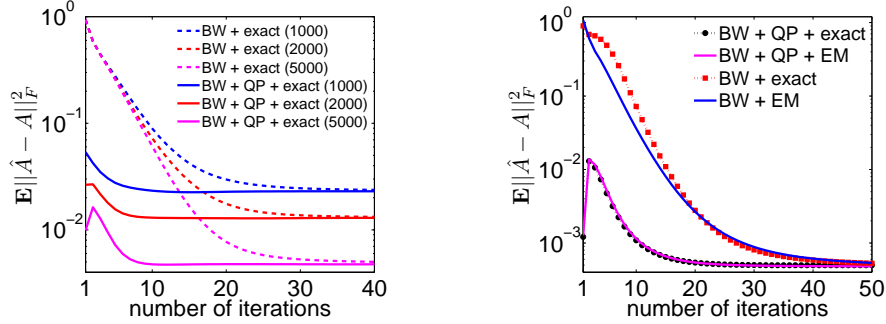


Figure 3: Convergence of the BW iterations.

the computation time. (iii) When the number of samples becomes large, the QP+EM is not only faster, but (surprisingly) also more accurate than BW+EM. As Fig. 3 suggests, this is due to the slow convergence of the BW algorithm, which requires more than 20 iterations for convergence. (iv) Starting the BW iterations with  $(\mu_i, \sigma_i^2)$  estimated by EM and  $A$  estimated by QP as its initial values significantly accelerated the convergence giving a superior accuracy after only 20 iterations. These results show the (well known) importance of initializing the BW algorithm with sufficiently accurate starting values. Our QP approach provides such an initial value for  $A$  by a computationally fast algorithm.

## References

- N. Abe and M.K. Warmuth. On the computational complexity of approximating distributions by probabilistic automata. *Machine Learning*, 9: 205–260, 1992.
- A. Anandkumar, D. Hsu, and S.M. Kakade. A method of moments for mixture models and hidden markov models. In *COLT*, 2012.
- L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41(1):pp. 164–171, 1970.
- M. Belkin and K. Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS)*, pages 103–112, 2010.
- P.J. Bickel, Y. Ritov, and T. Rydén. Asymptotic normality of the maximum-

- likelihood estimator for general hidden Markov models. *Ann. Statist.*, 26 (4):1614–1635, 1998.
- O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005.
- J.T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.
- G. Cybenko and V. Crespi. Learning hidden Markov models using nonnegative matrix factorization. *IEEE Trans. Information Theory*, 57(6):3963–3970, 2011.
- J.W. Daniel. Stability of the solution of definite quadratic programs. *Mathematical Programming*, 5:41–53, 1973.
- G.B. Dantzig, J. Folkman, and N. Shapiro. On the continuity of the minimum sets of a continuous function. *J. Math. Anal. Appl.*, 17:519–548, 1967.
- D. den Hertog. *Interior point approach to linear, quadratic and convex programming*, volume 277 of *Mathematics and its Applications*. Kluwer, Dordrecht, 1994.
- R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7(3):pp. 381–420, 2001.
- A. Faragó and G. Lugosi. An algorithm to find the global optimum of left-to-right hidden Markov model parameters. *Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform.*, 18(6):435–444, 1989.
- D. Hsu, S.M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *COLT*, 2009.
- A. Kontorovich and R. Weiss. Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes, arxiv:1207.4678. 2012.
- B. Lakshminarayanan and R. Raich. Non-negative matrix factorization for parameter estimation in hidden markov models. In *Machine Learning for Signal Processing (MLSP)*, pages 89–94, 2010.
- R. B. Lyngsø and C. N. Pedersen. Complexity of comparing hidden markov models. In *Proceedings of the 12th International Symposium on Algorithms and Computation*, pages 416–428. Springer-Verlag, 2001.

- M. Mohri and A. Rostamizadeh. Stability bounds for stationary  $\varphi$ -mixing and  $\beta$ -mixing processes. *The Journal of Machine Learning Research*, 11: 789–814, 2010.
- Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.
- E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. *Ann. Appl. Probab.*, 16(2):583–614, 2006.
- Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, Philadelphia, PA, 1994.
- L. R. Rabiner. Readings in speech recognition. chapter A tutorial on hidden Markov models and selected applications in speech recognition, pages 267–296. Morgan Kaufmann, 1990.
- S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Comput.*, 11:305–345, February 1999. ISSN 0899-7667.
- S. M. Siddiqi, B. Boots, and G. J. Gordon. Reduced-rank Hidden Markov Models. In *AISTAT*, 2010.
- S. Terwijn. On the learnability of Hidden Markov Models. In *Proceedings of the 6th International Colloquium on Grammatical Inference: Algorithms and Applications*, ICGI '02, pages 261–268, London, UK, 2002. Springer-Verlag.

## 6 Appendix

We now give a detailed account for the theorems stated in section 4.

### 6.1 Preliminaries I

In what follows we use the following notation: For an  $n \times n$  matrix  $A$ ,  $\text{vec}(A) \in \mathbb{R}^{n^2}$  is the result of stacking its columns vertically into a single long vector. Thus, its Frobenius matrix norm is  $\|A\|_F = \|\text{vec}(A)\|_2$ .

Recall the definition of  $g_\psi$ :

$$g_\psi \equiv \frac{2G}{1-\psi}.$$

One can easily verify that for  $2G \geq 1$ , we have  $1 + \psi g_\psi \leq g_\psi^2$ . Also recall that assumption (2b) states that the distributions in  $\mathcal{F}_n^\theta$  are bounded by  $L$ , which is defined by:

$$\max_{i \in [n]} \sup_{y \in \mathbb{R}} f_{\theta_i}(y) \leq L < \infty.$$

The following concentration result from Kontorovich and Weiss [2012, Theorem 1] is our main tool in proving the error bounds given here.

**Lemma 1.** *Let  $Y = Y_0, \dots, Y_{T-1} \in \mathcal{Y}^T$  be the output of a Hidden Markov chain with transition matrix  $A$  and output distributions  $\mathcal{F}_n^\theta$ . Assume that  $A$  is geometrically ergodic with constants  $G, \psi$  as in (1). Let  $F : (Y_0, \dots, Y_{T-1}) \mapsto \mathbb{R}$  be any function that is  $l$ -Lipschitz with respect to the Hamming metric on  $\mathcal{Y}^T$ . Then, for all  $\epsilon > 0$ ,*

$$P(|F(Y) - \mathbf{E}F| > \epsilon T) \leq 2 \exp\left(-\frac{T(1-\psi)^2 \epsilon^2}{2l^2 G^2}\right). \quad (32)$$

We will also need the following Lemma (proved in [Kontorovich and Weiss, 2012] for the discrete output case but easily generalize to continuous outputs) for bounding the variance of our estimators.

**Lemma 2.** *Let  $f(y) : \mathbb{R} \rightarrow \mathbb{R}^+$  be a function of the observables of an  $n$  states geometrically ergodic HMM with constants  $(G, \psi)$  and*

$$\int_{\mathcal{Y}} f(y) dy \leq 1.$$

*Assume the HMM is started with the stationary distribution  $\pi$ . Then*

$$\text{Var}\left[\frac{1}{T} \sum_{t=0}^{T-1} f(Y_t)\right] \leq \frac{\text{Var}[f(Y)]}{T} + \frac{\psi g_\psi \mathbf{E}[f(Y)]}{T}.$$

Similarly, let  $g(y, y') : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^+$  be a function of **consecutive** observations  $(y, y')$  such that

$$\iint_{\mathcal{Y}} g(y, y') dy dy' \leq 1.$$

Then

$$\begin{aligned} \text{Var} \left[ \frac{1}{T} \sum_{t=1}^{T-1} g(Y_t, Y_{t+1}) \right] &\leq \frac{\text{Var}[g(Y, Y')]}{T-1} + \\ &\quad \frac{(1 + \psi g_\psi) \mathbf{E}[g(Y, Y')]}{T-1}. \end{aligned}$$

## 6.2 Accuracy of $\hat{\rho}, \hat{\sigma}, \hat{\xi}$ and $\hat{\eta}$

Since our estimators  $\hat{\pi}$  and  $\hat{A}$  are constructed in terms of  $\hat{\rho}$  and  $\hat{\sigma}$  in the discrete case, and  $\hat{\xi}$  and  $\hat{\eta}$  in the continuous case, let us first examine the accuracy of the later. The following results shows that geometric ergodicity is sufficient to ensure their rapid convergence to the true values.

**Lemma 3. Discrete case.** *Let  $(y_t)_{t=1}^T$  be an observed sequence from a discrete output HMM whose initial state  $X_0$  follows the stationary distribution  $\pi$ . Let  $\rho$  be given by (3) and  $\sigma$  by (4) with their empirical estimates given in (5). Then*

$$\mathbf{E}[\|\hat{\rho} - \rho\|_2] \leq \sqrt{\frac{1 + \psi g_\psi}{T}} \quad (33)$$

$$\mathbf{E}[\|\hat{\sigma} - \sigma\|_2] \leq \sqrt{\frac{2 + \psi g_\psi}{T-1}} \quad (34)$$

Furthermore, for any  $\epsilon > 0$ ,

$$P(\|\hat{\rho} - \rho\|_2 > \sqrt{\frac{1 + \psi g_\psi}{T}} + \epsilon) \leq 2 \exp\left(-\frac{2T\epsilon^2}{g_\psi^2}\right) \quad (35)$$

and

$$\begin{aligned} P\left(\|\hat{\sigma} - \sigma\|_2 > \sqrt{\frac{2 + \psi g_\psi}{T-1}} + \epsilon\right) &\leq \\ &2 \exp\left(\frac{-2(T-1)\epsilon^2}{g_\psi^2}\right). \end{aligned} \quad (36)$$

Finally, we have for any fixed  $\mathbf{v} \in \mathbb{R}^m$  with  $\|\mathbf{v}\|_2 = 1$ ,

$$P(|\langle \hat{\boldsymbol{\rho}}, \mathbf{v} \rangle - \langle \boldsymbol{\rho}, \mathbf{v} \rangle| > \epsilon) \leq 2 \exp \left( -\frac{2T\epsilon^2}{g_\psi^2} \right). \quad (37)$$

*Proof.* First note that w.r.t the Hamming metric,  $T\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2$  and  $|\langle \hat{\boldsymbol{\rho}}, \mathbf{v} \rangle - \langle \boldsymbol{\rho}, \mathbf{v} \rangle|$  are 1-Lipschitz and  $T\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2$  is 2-Lipschitz. Thus the claims in (35, 36, 37) all follows directly from Lemma 1 where for (35, 36) we also take into account (33) and (34) respectively. In order to prove (33) note that

$$\mathbf{E}[\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2] = \sum_{k \in [n]} \mathbf{E}(\hat{\rho}_k - \rho_k)^2 = \sum_{k \in [n]} \text{Var}(\hat{\rho}_k).$$

So by taking in Lemma 2,  $f(y) = \mathbb{1}_{y=k}$ , we have  $\mathbf{E}[\mathbb{1}_{y=k}] = \rho_k$  and  $\text{Var}(\mathbb{1}_{y=k}) = \rho_k(1 - \rho_k) \leq \rho_k$ . Since  $\sum_{k=1}^m \rho_k = 1$  we get the desired bound.

The bound in (34) is obtained similarly by taking  $g(y, y') = \mathbb{1}_{y=k} \mathbb{1}_{y'=k'}$  in Lemma 2 with the fact that  $\sum_{kk'} \sigma_{kk'} = 1$ .  $\square$

**Lemma 4. Continuous case.** Let  $(Y_t)_{t=1}^T$  be an observed sequence from a continuous observations HMM whose initial state  $X_0$  follows the stationary distribution  $\boldsymbol{\pi}$ . Let  $\boldsymbol{\xi}$  be given by (13),  $\boldsymbol{\eta}$  by (18) and  $\hat{\boldsymbol{\xi}}$  and  $\hat{\boldsymbol{\eta}}$  be their empirical estimates, given by (14) and (19) respectively. Then for any  $\epsilon > 0$ ,

$$P \left( \left\| \hat{\boldsymbol{\xi}} - \boldsymbol{\xi} \right\|_2 > \epsilon \right) \leq 2n \exp \left( -\frac{2T\epsilon^2}{g_\psi^2 n L^2} \right), \quad (38)$$

and

$$P \left( \left\| \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \right\|_2 > \epsilon \right) \leq 2n^2 \exp \left( -\frac{2(T-1)\epsilon^2}{g_\psi^2 n^2} \right). \quad (39)$$

*Proof.* Note that  $\mathbf{E}\hat{\xi}_k = \xi_k$  and  $T\hat{\xi}_k$  is  $L$ -Lipschitz for all  $k \in [n]$ . Thus by Lemma 1 and the union bound we have

$$P \left( \left\| \hat{\boldsymbol{\xi}} - \boldsymbol{\xi} \right\|_\infty > \epsilon' \right) \leq 2n \exp \left( -\frac{2T\epsilon'^2}{g_\psi^2 L^2} \right). \quad (40)$$

Since

$$\left\| \hat{\boldsymbol{\xi}} - \boldsymbol{\xi} \right\|_2^2 = \sum_{k \in [n]} (\hat{\xi}_k - \xi_k)^2 \leq n \left\| \hat{\boldsymbol{\xi}} - \boldsymbol{\xi} \right\|_\infty^2,$$

we have

$$P\left(\left\|\hat{\xi} - \xi\right\|_2 > \epsilon\right) \leq P\left(\sqrt{n}\left\|\hat{\xi} - \xi\right\|_\infty > \epsilon\right).$$

putting  $\epsilon' = \epsilon/\sqrt{n}$  in (40), the claim in (38) follows.

The proof of (39) follows the same paradigm as the proof for (40). Indeed  $\mathbf{E}[\hat{\eta}_{kk'}] = \eta_{kk'}$  and  $T\eta_{kk'}$  is 1-Lipschitz so by Lemma 1 and the union bound we have

$$P\left(\|\hat{\eta} - \eta\|_\infty > \epsilon'\right) \leq 2n^2 \exp\left(-\frac{2T\epsilon'^2}{g_\psi^2 L^2}\right). \quad (41)$$

Since

$$\|\hat{\eta} - \eta\|_2^2 = \sum_{k,k' \in [n] \times [n]} (\hat{\eta}_{kk'} - \eta_{kk'})^2 \leq n^2 \|\hat{\eta} - \eta\|_\infty^2,$$

we have

$$P\left(\|\hat{\eta} - \eta\|_2 > \epsilon\right) \leq P\left(n\|\hat{\eta} - \eta\|_\infty > \epsilon\right).$$

putting  $\epsilon' = \epsilon/n$  in (41), the claim in (39) follows.  $\square$

### 6.3 Proof of theorem 1 - Strong consistency

We now prove the strong consistency of our estimators stated in Theorem 1.

*Proof.* For the discrete case, by Lemma 3, the expectation  $\mathbf{E}[\|\hat{\rho} - \rho\|_2]$  goes to zero as  $T \rightarrow \infty$ . Furthermore, using the Borel-Cantelli lemma,  $\|\hat{\rho} - \rho\|_2$  converge to its expectation a.s. concluding that  $\hat{\rho}$  converges a.s. to  $\rho$ . The same argument goes for  $\hat{\sigma}, \hat{\xi}, \hat{\eta}$  and  $\sigma, \xi, \eta$  respectively.

Now, the function  $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$  given by  $f(x) = (B^\top \text{diag}(1/x)B)^{-1}\mathbf{1}$  is continuous on  $\mathbb{R}_+^m$ . Moreover,  $f(\rho) = \pi$  since the optimization problem (7) has a unique minimizer  $x^*$  for all  $\hat{\rho}$ , which in particular is given by  $x^* = \pi$  when  $\hat{\rho} = \rho$ . Since  $\rho \in \mathbb{R}_+^m$  by assumption, the argument above shows that almost surely,  $\hat{\rho} \in \mathbb{R}_+^m$  for all sufficiently large  $T$ . Therefore,  $\lim_{T \rightarrow \infty} f(\hat{\rho}) = f(\rho) = \pi$  almost surely, and the asymptotic strong consistency of  $\hat{\pi}$  is established.

To prove the asymptotic strong consistency of  $\hat{A}$  in the discrete case, recall that the minimizer of the quadratic program  $x^\top Kx - h^\top x$  subject to  $Gx \leq g, Dx = d$ , is continuous under small perturbations of  $K, h, G, D, d$



[Dantzig et al., 1967]. In particular, if  $\hat{\pi}$  is sufficiently close to  $\pi$  then  $\hat{A}$  is close to  $A$ . Since  $\hat{\pi} \rightarrow \pi$  and  $\hat{\sigma} \rightarrow \sigma$  almost surely, we also have  $\hat{A} \xrightarrow{\text{a.s.}} A$ .

For the continuous observations case, note that  $\hat{\pi}$  and  $\hat{A}$  are also solutions of quadratic programs. Also note that  $\hat{\xi} \rightarrow \xi$  and  $\hat{\eta} \rightarrow \eta$  almost surely. Thus we have that  $\hat{A} \xrightarrow{\text{a.s.}} A$  and  $\hat{\pi} \xrightarrow{\text{a.s.}} \pi$  as above.  $\square$

#### 6.4 Proof of Theorem 2: Bounding the error for $\hat{\pi}$ in the discrete observations case

*Proof.* Lemma 3 and the fact that  $\|\hat{\rho} - \rho\|_\infty \leq \|\hat{\rho} - \rho\|_2$  implies that  $\|\hat{\rho} - \rho\|_\infty = O_P(1/\sqrt{T})$ . Hence we make a change of variables,

$$\hat{\rho} = \rho + \frac{1}{\sqrt{T}}\zeta. \quad (42)$$

To establish the (eventual) positivity of the entries of  $\hat{\pi}$ , we consider the solution  $x^*$  of (8) with  $\lambda = 0$ , e.g. without the normalization  $\sum x_i = 1$ , and write it as  $x^* = \pi + \delta$ . Our goal is to understand the relation between  $\delta$  and  $\zeta$ .

Observe that  $\delta$  satisfies the system of linear equations

$$\sum_j \left( \sum_k \frac{B_{kj}B_{ki}}{\rho_k \left(1 + \frac{1}{\sqrt{T}} \frac{\zeta_k}{\rho_k}\right)} \right) (\pi_j + \delta_j) = 1.$$

We need  $T$  sufficiently large so that, with high probability,  $\max_k \frac{1}{\sqrt{T}} \frac{\zeta_k}{\rho_k} \ll 1$ , or equivalently,  $|\hat{\rho}_k - \rho_k| \ll \rho_k$ .

By taking  $T \gtrsim 4g_\psi/a_1^2$  we have

$$\mathbf{E}[\|\hat{\rho} - \rho\|_\infty] \leq a_1/2.$$

So choosing  $\epsilon = \min \rho_k/2 \geq a_1/2$  in (35), this condition is satisfied for  $T \gtrsim g_\psi^2/a_1^2$ . Then, approximating  $1/(1 + \epsilon) = 1 - \epsilon + O(\epsilon^2)$  gives

$$\begin{aligned} \sum_j \left[ \sum_k \frac{B_{kj}B_{ki}}{\rho_k} \left( 1 - \frac{1}{\sqrt{T}} \frac{\zeta_k}{\rho_k} \right) \right] (\pi_j + \delta_j) \\ = 1 + O_P\left(\frac{1}{T}\right). \end{aligned}$$

Note that since  $B\pi = \rho$ , the leading order correction for  $\delta$  is simply

$$\delta = \frac{1}{\sqrt{T}}(\tilde{B}^\top \tilde{B})^{-1} \tilde{B}^\top \begin{pmatrix} \zeta \\ \rho \end{pmatrix} + O_P\left(\frac{1}{T}\right),$$

where the matrix  $\tilde{B} = \text{diag}(1/\sqrt{\rho})B$ .

Let  $\{\mathbf{u}_i\}$  and  $\{\mathbf{v}_i\}$  be the right and left singular vectors of  $\tilde{B}$  with non-zero singular values  $\sigma_i(\tilde{B})$ , where  $\sigma_1 \leq \sigma_2 \leq \dots \leq \sigma_n$ ; thus,  $\tilde{B}\mathbf{u}_i = \sigma_i\mathbf{v}_i$ . The fact that  $\tilde{B}$  also has  $n$  non-zero singular values follows from its definition combined with our Assumption 2d that  $B$  has rank  $n$ . Then

$$\tilde{B}^\top \tilde{B} = \sum_i \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^\top \quad (43)$$

and hence,

$$\delta = \frac{1}{\sqrt{T}} \sum_i \frac{1}{\sigma_i} \langle \frac{\zeta}{\rho}, \mathbf{v}_i \rangle \mathbf{u}_i + O_P\left(\frac{1}{T}\right) \quad (44)$$

For the solution  $x$  to have strictly positive coordinates we need that  $|\delta_j| < \pi_j$  for each of  $j = 1, \dots, n$ . Without loss of generality, assume that  $\pi_1 = \min_j \pi_j$  and analyze the worst-case setting. This occurs when the singular vector  $\mathbf{u}_1$  with smallest singular value coincides with the standard basis vector  $\mathbf{e}_1$ . Then,

$$|\delta_1| \leq \frac{1}{\sqrt{T}} \frac{1}{\sigma_1(\tilde{B}) \min_j \rho_j} |\langle \zeta, \mathbf{v}_1 \rangle| + O_P\left(\frac{1}{T}\right). \quad (45)$$

It follows from (37) that  $|\delta_1|$  will be dominated by  $\min \pi_j \geq a_0$  provided that

$$T \gtrsim \frac{g_\psi}{a_0 a_1 \sigma_1(\tilde{B})}. \quad (46)$$

In the unlikely event that (i) the vector  $\boldsymbol{\pi}$  is uniform ( $\pi_j = 1/n$  for all  $j$ ), (ii) the matrix  $\tilde{B}$  has  $n$  identical singular values, we need the equation analogous to (45) to hold for all  $n$  coordinates. By a union bound argument, an additional factor of  $\log n$  in the number of samples suffices to ensure, with high probability, the non-negativity of the solution  $x$ .

Next we proceed to bound  $\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2^2$ . To this end, we write

$$x^* - \boldsymbol{\pi} = \delta = \sum_i \frac{1}{\sigma_i(\tilde{B})} \langle \frac{\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}}{\boldsymbol{\rho}}, \mathbf{v}_i \rangle \mathbf{u}_i + O_P\left(\frac{1}{T}\right).$$

Since both the  $\{\mathbf{u}_i\}$  and the  $\{\mathbf{v}_i\}$  are orthonormal,

$$\begin{aligned} \|\delta\|_2^2 &= \sum_i \frac{1}{\sigma_i^2(\tilde{B})} \langle \frac{\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}}{\boldsymbol{\rho}}, \mathbf{v}_i \rangle^2 \\ &\leq \frac{1}{\sigma_1^2(\tilde{B}) (\min \rho_k)^2} \sum_i \langle \hat{\boldsymbol{\rho}} - \boldsymbol{\rho}, \mathbf{v}_i \rangle^2 \\ &\leq \frac{\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2}{\sigma_1^2(\tilde{B}) a_1^2}. \end{aligned}$$

Bounding  $\|\hat{\rho} - \rho\|_2^2$  via Lemma 3 and noting that

$$\|\hat{\pi} - \pi\|_2 = \left\| \frac{x^*}{\|x^*\|_1} - \pi \right\|_2 \leq 2 \|x^* - \pi\|_2 = 2 \|\delta\|_2,$$

the result in (22) follows.  $\square$

## 6.5 Preliminaries II

The remaining estimators ( $\hat{\pi}$  for the continuous observations case, and  $\hat{A}$  for both the discrete and continuous observations cases) are obtained as solutions for quadratic programs. Let us take for example the QP for calculating  $\hat{\pi}$  with continuous observations HMM, given in (23). For this case, the QP is equivalent to

$$\hat{\pi} = \operatorname{argmin}_x \frac{1}{2} x^\top K^\top K x - x^\top K^\top \hat{\xi}$$

subject to  $x \geq 0$  and  $\sum_i x_i = 1$ .

Note that if  $\hat{\xi}$  was equal to its true values  $\xi$ , the solution of the above QP would simply be the true  $\pi$ . In reality, we only have the estimate  $\hat{\xi}$ . In order to analyze the error  $\|\hat{\pi} - \pi\|_2$ , we will need to consider how the solutions of such a quadratic program are affected by errors in  $\xi$ .

More generally, we are concerned with two QPs

$$\min Q(x) = \min \frac{1}{2} x^\top M x - x^\top h, \quad (47)$$

$$\min \hat{Q}(x) = \min \frac{1}{2} x^\top \hat{M} x - x^\top \hat{h}, \quad (48)$$

both subject to  $Gx \leq g$ ,  $Dx = d$ . We assume that the solution to the first QP is the “true” value while the solution to the second is our estimate. So bounding the estimate error is equivalent to bounding the error between the solutions obtained by the above two QPs, where  $\hat{M}$  and  $\hat{h}$  are perturbed versions of  $M$  and  $h$ .

Given that, note that only the objective function has been perturbed, while the linear constraints remained unaffected. We may thus apply the following classical result on the solution stability of definite quadratic programs.

**Theorem 7.** *[Daniel, 1973] Let  $\lambda = \lambda_{\min}(M)$  be the smallest eigenvalue of  $M$ , and let  $\epsilon = \max\{\|\hat{M} - M\|_2, \|\hat{h} - h\|_2\}$ . Let  $x$  and  $\hat{x}$  be the minimizers of Eqs.(47) and (48), respectively. Then, for  $\epsilon < \lambda$ ,*

$$\|x - \hat{x}\|_2 \leq \frac{\epsilon}{\lambda - \epsilon} (1 + \|x\|_2).$$

In the following we will obtain bounds on  $\epsilon$  and  $\lambda$  for the different estimators and invoke the above theorem.

### 6.6 Proof of Theorem 3: Bounding the error for $\hat{\pi}$ in the continuous observations case

*Proof.* Note that in the notation given in Theorem 7, we have  $h = \xi^\top K$  and  $\hat{h} = \hat{\xi}^\top K$ . Since we assumed that the output density parameters are known exactly we have no error in  $M = K^\top K$ .

It is immediate that

$$\lambda_{\min}(K^\top K) = \sigma_1^2(K),$$

and

$$\epsilon \leq \|\hat{\xi} - \xi\|_2 \|K\|_2 \leq nL \|\hat{\xi} - \xi\|_2.$$

From Lemma 4 we have

$$\|\hat{\xi} - \xi\|_2 \lesssim_P \sqrt{\frac{(n \ln n) g_\psi^2 L^2}{T}},$$

while by Theorem 7 we have

$$\|\hat{\pi} - \pi\|_2 \lesssim \frac{\epsilon}{\lambda_{\min}(K^\top K)} (1 + \|\pi\|_2).$$

Since  $\|\pi\|_2 \leq 1$ , the claim follows.  $\square$

As a side remark we note that the form of (24) is somewhat counter-intuitive, as it suggests a worse behavior for larger  $L$ . Intuitively, however, larger  $L$  corresponds to a more peaked — and hence lower-variance — density, which ought to imply sharper estimates. Note however that as numerical simulations suggest we typically have

$$\frac{\sigma_1^2(\tilde{F}) L^2}{\sigma_1^2(\tilde{K})} = O(1).$$

Thus, whenever  $\sigma_1^2(\tilde{F})$  is well behaved so is the estimate in (24) and the bound is reasonable after all. Finally note that  $F$  is stochastic so it behaves very much like the matrix  $B$  in the discrete outputs case.

## 6.7 Proof of Theorem 4: Bounding the error of $\hat{A}$ in the discrete observations case

Let  $\hat{A}$  be the solution of

$$\min_{A_{ij} \geq 0, \sum_i A_{ij} = 1} \|\hat{\sigma} - \hat{C}A\|_2^2, \quad (25)$$

where  $\hat{\sigma}$  is given in (5). Recall that  $C_{ij}^{kk'} = \pi_j B_{kj} B_{k'i}$  and  $\hat{C}_{ij}^{kk'} = \hat{\pi}_j B_{kj} B_{k'i}$ . First note that if  $\pi$  and  $\sigma$  were known exactly, the above QP could be written as

$$\min Q(A) = \min \frac{1}{2} \text{vec}(A)^\top M \text{vec}(A) - \text{vec}(A)^\top h \quad (49)$$

where  $M = C^\top C$  and  $h = C^\top \text{vec}(\sigma)$ . Its solution is precisely the transition probability matrix  $A$ . In reality, as we only have estimates  $\hat{\pi}$  and  $\hat{\sigma}$ , the optimization problem is perturbed to

$$\min \hat{Q}(A) = \min \frac{1}{2} \text{vec}(A)^\top \hat{M} \text{vec}(A) - \text{vec}(A)^\top \hat{h} \quad (50)$$

where  $\hat{M} = \hat{C}^\top \hat{C}$ , and  $\hat{h} = \hat{C}^\top \text{vec}(\hat{\sigma})$ .

To analyze how errors in  $\hat{\sigma}$  and  $\hat{C}$  affect the optimization problem we follow the same route as above. Thus we need to bound  $\|\hat{h} - h\|_2$ ,  $\|\hat{M} - M\|_2$ , and the smallest eigenvalue of  $M$ . Regarding the latter, by definition,  $\lambda_{\min}(M) = \sigma_1^2(C)$ , where  $\sigma_1(C)$  is the smallest singular value of  $C$ . A simple exercise in linear algebra yields

$$\sigma_1(C) \geq a_0 \sigma_1^2(B). \quad (51)$$

The following lemma provides bounds on  $\|\hat{M} - M\|_2$  and on  $\|\hat{h} - h\|_2$ .

**Lemma 5.** *Asymptotically, as  $T \rightarrow \infty$ ,*

$$\|\hat{h} - h\|_2 \lesssim_P \sqrt{n} (\|\hat{\pi} - \pi\|_2 + \|\hat{\sigma} - \sigma\|_2) \quad (52)$$

and

$$\|\hat{M} - M\|_2 \lesssim_P 2n \|\hat{\pi} - \pi\|_2. \quad (53)$$

*Proof.* By definition,  $h_{ij} = \sum_{k,k'} C_{ij}^{kk'} \sigma_{kk'}$ , and  $\hat{h}_{ij} = \sum_{k,k'} \hat{C}_{ij}^{kk'} \hat{\sigma}_{kk'}$ . Using the definitions of  $C$  and  $\hat{C}$ , up to mixed terms  $O(\|\hat{\pi} - \pi\|_\infty \|\hat{\sigma} - \sigma\|_\infty)$ , we obtain

$$\begin{aligned} \hat{h}_{ij} - h_{ij} &= (\hat{\pi}_j - \pi_j) \sum_{kk'} B_{kj} B_{k'i} \sigma_{kk'} \\ &\quad + \pi_j \sum_{kk'} B_{kj} B_{k'i} (\hat{\sigma}_{kk'} - \sigma_{kk'}) \end{aligned}$$

Since each of  $\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty$  and  $\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_\infty$  are  $O_P(1/\sqrt{T})$ , the neglected mixed terms are asymptotically negligible as compared to each of the first two ones. Next, we use the fact that  $\sigma_{kk'} \leq 1$ ,  $\pi_j \leq 1$  and  $\sum_{kk'} B_{kj} B_{k'i} \leq 1$  to obtain that

$$\left\| \hat{h} - h \right\|_2 \lesssim_P \sqrt{n} \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 + \sqrt{n} \|\text{vec}(\hat{\boldsymbol{\sigma}}) - \text{vec}(\boldsymbol{\sigma})\|_2$$

Similarly, we have that for the  $n^2 \times n^2$  matrix  $M$ , and not including higher order mixed terms  $(\hat{\pi}_j - \pi_j)(\hat{\pi}_\beta - \pi_\beta)$ , which are asymptotically negligible,

$$\begin{aligned} (\hat{M} - M)_{ij,\alpha\beta} &= (\hat{\pi}_j - \pi_j) \pi_\beta \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} \\ &\quad + (\hat{\pi}_\beta - \pi_\beta) \pi_j \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} \end{aligned}$$

Note that  $\sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} = (\sum_k B_{kj} B_{k\beta})(\sum_{k'} B_{k'i} B_{k'\alpha}) \leq 1$ . Hence, by similar arguments as for  $h$ , (53) follows.  $\square$

We can now prove Theorem 4:

*Proof. (of Theorem 4)* Lemma 3, together with (22), implies that with high probability,

$$\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_F \lesssim_P \sqrt{\frac{g_\psi^2}{T-1}},$$

and

$$\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 \lesssim_P \sqrt{\frac{g_\psi^2}{T a_1^2 \sigma_1^2(\tilde{B})}}.$$

Inserting these into (52) and (53) yields, w.h.p.,

$$\begin{aligned} \epsilon &= \max \left\{ \left\| \hat{h} - h \right\|_2, \left\| \hat{M} - M \right\|_2 \right\} \\ &\lesssim \sqrt{\frac{n^2 g_\psi^2}{T a_1^2 \sigma_1^2(\tilde{B})}}. \end{aligned} \tag{54}$$

By Theorem 7, we have that

$$\left\| \hat{A} - A \right\|_F \lesssim \frac{\epsilon}{\lambda_1(M)} (1 + \|A\|_F), \tag{55}$$

where  $\|A\|_F \leq \sqrt{n}$  since  $A$  is column-stochastic. The claim follows by substituting the bounds on  $\epsilon$  in (54) and on  $\lambda_1(M) = \sigma_1^2(C) \geq a_0^2 \sigma_1^4(B)$  in (51) into (55) and noting that  $\sigma_1^2(\tilde{B}) \geq \sigma_1^2(B)$ .  $\square$

## 6.8 Proof of Theorem 5: Bounding the error of $\hat{A}$ in the continuous observations case

Let  $\hat{A}$  be the solution of

$$\min_{A_{ij} \geq 0, \sum_i A_{ij} = 1} \|\hat{\boldsymbol{\eta}} - \hat{C}A\|_2^2, \quad (25)$$

where  $\hat{\boldsymbol{\eta}}$  is given in (19) and  $C_{ij}^{kk'} = \pi_j F_{kj} F_{k'i}$  and  $\hat{C}_{ij}^{kk'} = \hat{\pi}_j F_{kj} F_{k'i}$ . The above QP can be written as

$$\min \hat{Q}(A) = \min \frac{1}{2} \text{vec}(A)^\top \hat{M} \text{vec}(A) - \text{vec}(A)^\top \hat{h} \quad (56)$$

where  $\hat{M} = \hat{C}^\top \hat{C}$ , and  $\hat{h} = \hat{C}^\top \text{vec}(\hat{\boldsymbol{\sigma}})$ .

Exactly as in the previous subsection, we want to bound the difference between the solutions for the above QP and the unperturbed one.

First note that

$$\sigma_1(C) \geq a_0 \sigma_1^2(F). \quad (57)$$

Next we give the analogue of lemma 5.

**Lemma 6.** *Asymptotically, as  $T \rightarrow \infty$ ,*

$$\|\hat{h} - h\|_2 \lesssim_P \sqrt{n} \left( \frac{1}{a_0} \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 + \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 \right) \quad (58)$$

and

$$\|\hat{M} - M\|_2 \lesssim_P 2n \frac{\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2}{a_0}. \quad (59)$$

*Proof.* In contrast to Lemma 5, here  $F$  is also perturbed due to errors in  $\hat{\boldsymbol{\pi}}$  with

$$\hat{F}_{ij} = \int_{\mathcal{Y}} \frac{\hat{\pi}_i f_i(y) f_j(y)}{\sum_k \hat{\pi}_k f_k(y)} dy.$$

Expanding the difference  $\Delta F_{ij} \equiv \left| \hat{F}_{ij} - F_{ij} \right|$  up to first order in  $\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}$  we find that

$$\|\Delta F\|_F \leq \frac{\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty}{a_0} \|F\|_F \leq \frac{\sqrt{n} \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty}{a_0},$$

where in the last inequality we used the fact that  $F$  is stochastic. Repeating the arguments in the proof for Lemma 5 and noting that  $a_0 \ll 1$  we get (58) and (59).  $\square$

We now come to the proof of Theorem 5.

*Proof. (of Theorem 5)* Lemma 4, together with (24), implies that with high probability,

$$\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_F \lesssim_P \sqrt{\frac{(n^2 \ln n) g_\psi^2}{T-1}},$$

and

$$\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 \lesssim_P \sqrt{\frac{(n^3 \ln n) g_\psi^2 L^4}{T \sigma_1^4(\tilde{K})}}$$

Inserting these into (58) and (59) yields, w.h.p.,

$$\begin{aligned} \epsilon &= \max \left\{ \left\| \hat{h} - h \right\|_2, \left\| \hat{M} - M \right\|_2 \right\} \\ &\lesssim \sqrt{\frac{(n^5 \ln n) g_\psi^2 L^4}{T \sigma_1^4(\tilde{K})}}. \end{aligned} \quad (60)$$

By Theorem 7, we have that

$$\left\| \hat{A} - A \right\|_F \lesssim \frac{\epsilon}{\lambda_1(M)} (1 + \|A\|_F), \quad (61)$$

where  $\|A\|_F \leq \sqrt{n}$  since  $A$  is column-stochastic. The claim follows by substituting the bounds on  $\epsilon$  in (60) and on  $\lambda_1(M) = \sigma_1^2(C) \geq a_0^2 \sigma_1^4(F)$  in (51) into (61) and noting that  $\sigma_1^2(\tilde{F}) \geq \sigma_1^2(F)$ .  $\square$

As for remark 1, we point out that estimating  $\boldsymbol{\eta}'$  with the help of the matrix  $K$  (instead of  $\boldsymbol{\eta}$  with  $F$ ) results in an estimator that is not  $O(1/T)$ -Lipschitz any more but  $O(L^2/T)$ -Lipschitz with  $L = \max_{i \in [n]} \sup_{y \in \mathbb{R}} f_{\theta_i}(y)$ . This means that in principle we will need many more samples to accurately estimate  $\boldsymbol{\eta}'$  compared to  $\boldsymbol{\eta}$ , see Lemma 4. Thus, since in high dimensions calculating  $F$  via numerical integration may be computational intensive, choosing between the two estimators is in some sense choosing between working with limited number of samples and computational efficiency.

## 6.9 Proof of Theorem 6: Perturbations in the output parameters

We give here the proof for the perturbation in the matrix  $F$ . The proof for perturbations in the matrix  $K$  is similar.



*Proof.* By definition,  $b_{ij} = \sum_{k,k'} C_{ij}^{kk'} \sigma_{kk'}$ , and  $\hat{b}_{ij} = \sum_{k,k'} \hat{C}_{ij}^{kk'} \hat{\sigma}_{kk'}$ . Using the definitions of  $C$  and  $\hat{C}$ , up to first order in  $\{\|\hat{\pi} - \pi\|_\infty, \|\hat{\sigma} - \sigma\|_\infty, \epsilon_F\}$  we obtain

$$\begin{aligned} \hat{b}_{ij} - b_{ij} &= (\hat{\pi}_j - \pi_j) \sum_{kk'} B_{kj} B_{k'i} \sigma_{kk'} \\ &\quad + \pi_j \sum_{kk'} B_{kj} B_{k'i} (\hat{\sigma}_{kk'} - \sigma_{kk'}) \\ &\quad + \epsilon_F \pi_j \sum_{kk'} (P_{kj} B_{k'i} + B_{kj} P_{k'i}) \sigma_{kk'}. \end{aligned}$$

As the two first terms already considered we focus on the last term. It can be shown that:

$$\sum_{ij} \left( \pi_j \sum_{kk'} P_{kj} B_{k'i} \sigma_{kk'} \right)^2 \leq n \|P\|_F^2.$$

Thus

$$\begin{aligned} \|\hat{b} - b\|_2 &\leq \sqrt{n} (\|\hat{\pi} - \pi\|_2 + \|vec(\hat{\sigma}) - vec(\sigma)\|_2 + \\ &\quad + 2\epsilon_F \|P\|_F) (1 + o(1)). \end{aligned} \tag{62}$$

Similarly, for the matrix  $K$  up to first order in  $\{\|\hat{\pi} - \pi\|_\infty, \epsilon_F\}$  we have

$$\begin{aligned} (\hat{K} - K)_{ij,\alpha\beta} &= (\hat{\pi}_j - \pi_j) \pi_\beta \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} \\ &\quad + (\hat{\pi}_\beta - \pi_\beta) \pi_j \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} \\ &\quad + \epsilon_F \pi_j \pi_\beta \sum_{kk'} P_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} + \dots \\ &\quad + \epsilon_F \pi_\beta \pi_j \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} P_{k'\alpha}. \end{aligned}$$

Again considering only the terms including  $P$  and using the facts that  $\sum_k B_{kj} B_{k\beta} \leq 1$  and  $\sum_{kk'} (P_{kj} B_{k'i})^2 \leq \sum_k P_{kj}^2$  we similarly find that

$$\|\hat{K} - K\|_2 \leq (1 + o_p(1)) 2n (\|\hat{\pi} - \pi\|_2 + 4\epsilon_F \|P\|_F).$$

Repeating the analysis in the proofs for Theorems 3, 4 and 5 give the desired result.  $\square$